

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Utility Patent Application

ROBUST BAYESIAN MIXTURE MODELING

Inventor(s):

Christopher M. Bishop

Markus Svensén

CLIENT'S DOCKET NO. MS305414.01

ATTORNEY'S DOCKET NO. MS1-1673US

EV3 95541585

ROBUST BAYESIAN MIXTURE MODELING

Technical Field

The invention relates generally to statistical analysis and machine learning algorithms, and more particularly to robust Bayesian mixture modeling.

Background

Mixture models are common tools of statistical analysis and machine learning. For example, when trying to model a statistical data distribution, a single Gaussian model may not adequately approximate the data, particularly when the data has multiple modes or clusters (e.g., has more than one peak).

As such, a common approach is to use a mixture of two or more Gaussian components, fitted with a maximum likelihood, to model such data. Nevertheless, even a mixture of Gaussians (MOG) presents modeling problems, such as inadequate modeling of outliers and severe overfitting. For example, there are singularities in the likelihood function arising from the collapse of components onto individual data points – a pathological result.

Some problems with a pure MOG can be elegantly addressed by adopting a Bayesian framework to marginalize over the model parameters with respect to appropriate priors. The resulting Bayesian model likelihood can then be maximized with respect to the number of Gaussian components in the mixture, if the goal is model selection, or combined with a prior over the number of the components, if the goal is model averaging. One benefit to a Bayesian approach using a mixture of Gaussians is the elimination of maximum likelihood singularities, although it still lacks robustness to outliers. In addition, in the

1 Bayesian model selection context, the presence of outliers or other departures from
2 the empirical distribution of Gaussianity can lead to errors in the determination of
3 the number of clusters in the data.

4 Summary

5
6 Implementations described and claimed herein address the foregoing
7 problems using a Bayesian treatment of mixture models based on individual
8 components having Student distributions, which have heavier tails compared to
9 the exponentially decaying tails of Gaussians. The mixture of Student distribution
10 components is characterized by a set of modeling parameters. Tractable
11 approximations of the posterior distributions of individual modeling parameters
12 are optimized and used to generate a data model for a set of input data.

13 In some implementations, articles of manufacture are provided as computer
14 program products. One implementation of a computer program product provides a
15 computer program storage medium readable by a computer system and encoding a
16 computer program. Another implementation of a computer program product may
17 be provided in a computer data signal embodied in a carrier wave by a computing
18 system and encoding the computer program.

19 The computer program product encodes a computer program for executing
20 a computer process on a computer system. A modeling parameter is selected from
21 a plurality of modeling parameters characterizing a mixture of Student distribution
22 components. A tractable approximation of a posterior distribution for the selected
23 modeling parameter is computed based on an input set of data and a current
24 estimate of a posterior distribution of at least one unselected modeling parameter
25 in the plurality of modeling parameters. A lower bound of a log marginal

1 likelihood is computed as a function of current estimates of the posterior
2 distributions of the modeling parameters. The current estimates of the posterior
3 distributions of the modeling parameters include the computed tractable
4 approximation of the posterior distribution of the selected modeling parameter. A
5 probability density that models the input set of data is generated, if the lower
6 bound is satisfactorily optimized. The probability density includes the mixture of
7 Student distribution components, which is characterized by the current estimates
8 of the posterior distributions of the modeling parameters.

9 In another implementation, a method is provided. A modeling parameter is
10 selected from a plurality of modeling parameters characterizing a mixture of
11 Student distribution components. A tractable approximation of a posterior
12 distribution for the selected modeling parameter is computed based on an input set
13 of data and a current estimate of a posterior distribution of at least one unselected
14 modeling parameter in the plurality of modeling parameters. A lower bound of a
15 log marginal likelihood is computed as a function of current estimates of the
16 posterior distributions of the modeling parameters. The current estimates of the
17 posterior distributions of the modeling parameters include the computed tractable
18 approximation of the posterior distribution of the selected modeling parameter. A
19 probability density that models the input set of data is generated, if the lower
20 bound is satisfactorily optimized. The probability density includes the mixture of
21 Student distribution components, which is characterized by the current estimates
22 of the posterior distributions of the modeling parameters.

23 In another implementation, a system is provided. A tractable
24 approximation module computes a tractable approximation of a posterior
25

1 distribution for the selected modeling parameter based on an input set of data and
2 a current estimate of a posterior distribution of at least one unselected modeling
3 parameter in the plurality of modeling parameters. A lower bound optimizer
4 module computes a lower bound of a log marginal likelihood as a function of
5 current estimates of the posterior distributions of the modeling parameters. The
6 current estimates of the posterior distributions of the modeling parameters include
7 the computed tractable approximation of the posterior distribution of the selected
8 modeling parameter. A data model generator generates a probability density
9 modeling the input set of data, if the lower bound is satisfactorily optimized. The
10 probability density includes the mixture of Student distribution components. The
11 mixture of Student distribution components is characterized by the current
12 estimates of the posterior distributions of the modeling parameters.

13 Other implementations are also described and recited herein.

14 **Brief Descriptions of the Drawings**

15
16 FIG. 1 illustrates exemplary probability distributions for modeling a data
17 set.

18 FIG. 2 illustrates exemplary operations for robust Bayesian mixture
19 modeling.

20 FIG. 3 illustrates an exemplary robust Bayesian mixture modeling system.

21 FIG. 4 illustrates a system useful for implementing an embodiment of the
22 present invention.
23
24
25

Detailed Description

FIG. 1 illustrates exemplary probability distributions 100 for modeling a data set. A single Gaussian distribution 102 models an input data set of independent identically distributed (idd) data 104. Note that the mean 106 of the single Gaussian distribution 102 is pulled substantially to the right in order to accommodate the outlier data element 106, thereby compromising the accuracy of the Gaussian model as it applies to the given data set 104. In addition, the standard deviation of the distribution 102 is undesirably increased by the outlier 106.

In order to improve the modeling of the data 104, a mixture of Gaussian distributions 108 may be used. However, fitting the mixture 108 to the data set 104 using a maximum likelihood approach does not yield a usable optimal number of components because the maximum likelihood approach favors an ever more complex model, leading to the undesirable extreme of individual, infinite magnitude Gaussian distribution component for individual data point. While overfitting of Gaussian mixture models can be addressed to some extent using Bayesian inference, even then, Gaussian mixture models continue to lack robustness as to outliers.

A mixture of Student distributions 110 can demonstrate a significant improvement in robustness as compared to a mixture of Gaussian distributions. However, there is no closed form solution for maximizing the likelihood under a Student distribution. Furthermore, the maximum likelihood approach does not address the problem of overfitting. Therefore, a mixture of Student distributions 110 combined with a tractable Bayesian treatment to fit the Student

1 mixture to the input data 104 addresses these issues, as illustrated in FIG. 1.
 2 However, no satisfactory method or system for obtaining a tractable Bayesian
 3 treatment of Student mixture distributions has previously been demonstrated. As
 4 such, in one implementation, robust Bayesian mixture modeling obtains a tractable
 5 Bayesian treatment of Student mixture distributions based on variational inference.
 6 In another implementation, a tractable approximation may be obtained using
 7 Monte Carlo-based techniques.

8 Robust Bayesian mixture modeling is based on a mixture of component
 9 distributions given by a multivariate Student distribution, also known as a t -
 10 distribution. A Student distribution represents a generalization of a Gaussian
 11 distribution and, in the limit $\nu \rightarrow \infty$, the Student distribution reduces to a Gaussian
 12 distribution with mean μ and precision Λ (i.e., inverse covariance). For finite
 13 values of ν , the Student distribution has heavier tails than the corresponding
 14 Gaussian having the same μ and Λ .

15 A Student distribution over a d -dimensional random variable \mathbf{x} may be
 16 represented in the following form:

$$17 \quad S(\mathbf{x}|\mu, \Lambda, \nu) = \frac{\Gamma\left(\frac{\nu+d}{2}\right) |\Lambda|^{\frac{1}{2}}}{\Gamma\left(\frac{\nu}{2}\right) (\nu\pi)^{\frac{d}{2}}} \left(\frac{\Delta^2}{\nu} + 1\right)^{-\frac{\nu+d}{2}} \quad (1)$$

18 where $\Delta^2 = (\mathbf{x} - \mu)^\top \Lambda (\mathbf{x} - \mu)$ represents the squared Mahalanobis distance from \mathbf{x}
 19 to μ .
 20
 21

22 In contrast to the Gaussian distribution, no closed form solution for
 23 maximizing likelihood exists under a Student distribution. However, the Student
 24 distribution may be represented as an infinite mixture of scaled Gaussian
 25

distributions over \mathbf{x} with an additional random variable u , which acts as a scaling parameter of the precision matrix Λ , such that the Student distribution may be represented in the following form:

$$S(\mathbf{x}|\boldsymbol{\mu}, \Lambda, \nu) = \int_0^\infty N(\mathbf{x}|\boldsymbol{\mu}, \Lambda u) G\left(u|\frac{\nu}{2}, \frac{\nu}{2}\right) du \quad (2)$$

where $N(\mathbf{x}|\boldsymbol{\mu}, \Lambda)$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and precision matrix Λ , and $G(u|a, b)$ represents the Gamma distribution. For each observation of \mathbf{x} (i.e., of N observations), a corresponding implicit posterior distribution over the variable u exists.

The probability density of mixtures of M Student distributions may be represented in the form:

$$p(\mathbf{x}|\{\boldsymbol{\mu}_m, \Lambda_m, \nu_m\}, \boldsymbol{\pi}) = \sum_{m=1}^M \pi_m S(\mathbf{x}|\boldsymbol{\mu}_m, \Lambda_m, \nu_m) \quad (3)$$

where the mixing coefficients $\boldsymbol{\pi} = (\pi_1, \dots, \pi_M)^\top$ satisfy $0 \leq \pi_m \leq 1$ and $\sum_{m=1}^M \pi_m = 1$.

In order to find a tractable treatment of this model, the mixture density of Equation (3) may be expressed in terms of a marginalization over a binary latent labeling variable \mathbf{s} of dimensions $N \times M$ (i.e., N representing the number of data elements and M representing the number of Student distribution components in the mixture) and the unobserved variable u_{nm} , also of dimensions $N \times M$ when applied to a mixture. Variable \mathbf{s} has components $\{s_{nj}\}$, such that $s_{nm}=1$ and $s_{nj}=0$ for $j \neq m$, resulting in:

$$p(\mathbf{x}_n|\mathbf{s}, \{\boldsymbol{\mu}_m, \Lambda_m, \nu_m\}) = \prod_{n,m} S(\mathbf{x}_n|\boldsymbol{\mu}_m, \Lambda_m, \nu_m)^{s_{nm}} \quad (4)$$

with a corresponding prior distribution over \mathbf{s} of the form:

$$p(\mathbf{s}|\boldsymbol{\pi}) = \prod_{n,m} \pi_m^{s_{nm}} \quad (5)$$

It can be verified that marginalization of the product of Equations (4) and (5) over the latent variable \mathbf{s} recovers the Student distribution mixture of Equation (3).

An input data set \mathbf{X} includes N iid observations \mathbf{x}_n , where $n=1, \dots, N$, which are assumed to be drawn independently from the distribution characterized by Equation (3). Thus, for each data observation \mathbf{x}_n , a corresponding discrete latent variable \mathbf{s}_n specifies which component of the mixture generated that data point, and continuous latent variable \mathbf{u}_{nm} specifies the scaling of the precision for the corresponding equivalent Gaussian distribution from which the data was hypothetically generated.

In addition to the prior distribution over \mathbf{s} , prior distributions for the modeling parameters $\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m$, and $\boldsymbol{\pi}$, are used in a Bayesian treatment of probability density estimation. As such, distributions of the modeling parameters are used rather than the parameters themselves. In one implementation, for tractability, conjugate priors from the exponential family have been chosen in the form:

$$p(\boldsymbol{\mu}_m) = N(\boldsymbol{\mu}_m | \mathbf{m}, \rho \mathbf{I}) \quad (6)$$

$$p(\boldsymbol{\Lambda}_m) = W(\boldsymbol{\Lambda}_m | \mathbf{W}_0, \eta_0) \quad (7)$$

$$p(\boldsymbol{\pi}) = D(\boldsymbol{\pi} | \boldsymbol{\alpha}) \quad (8)$$

wherein $W(\Lambda|\square\square)$ represents the Wishart distribution and $D(\pi|\square)$ represents the Dirichlet distribution. The prior $p(u)$ is implicitly defined in Equation (2) to equal the Gamma distribution $G\left(u|\frac{\nu}{2}, \frac{\nu}{2}\right)$.

It should be understood that prior distributions may be selected from other members of the exponential family in alternative embodiments. The parameters of the prior distributions on μ and Λ are chosen to give broad distributions (e.g., in one implementation, $m_0=0$, $\rho_0=10^{-3}$, $W_0=I$, $\eta_0=1$. For the prior distribution over π , $\alpha=\{\alpha_m\}$ are interpreted as effective numbers of prior observations, with $\alpha_m=10^{-3}$.

Exact inference of the Bayesian model is intractable. However, with the choice of exponential distributions to represent the prior distributions of the modeling parameters, tractable approximations are possible. In one implementation, for example, a tractable approximation may be obtained through Monte Carlo techniques.

In another implementation, variational inference may be employed to obtain tractable approximations of the posterior distributions over the identified stochastic modeling parameters, which in one implementation includes $\{\mu_m, \Lambda_m\}$, π , and $\{s_m, u_n\}$. (Another modeling parameter, ν , is treated in a deterministic (i.e., non-stochastic) fashion; however, only one such parameter exists per mixture component.).

In variational inference, the log-marginal likelihood is maximized. One form of the log-marginal likelihood is shown:

$$\ln \prod_n^N p(x_n | m_0, \rho_0, W_0, \eta_0) = \ln \int \prod_n^N p(x_n, u_n | \mu, \Lambda, \nu) p(\mu | m_0, \rho_0) p(\Lambda | W, \eta) du_n d\mu d\Lambda \quad (9)$$

This quantity cannot be maximized directly. However, Equation (9) can be re-written as follows:

$$\begin{aligned} & \ln \int p(\mathbf{X}|\theta) p(\theta|m_0, \rho_0, \mathbf{W}_0, \eta_0, \nu) d\theta \\ &= \int q(\theta) \ln \frac{p(\mathbf{X}, \theta|m_0, \rho_0, \mathbf{W}_0, \eta_0, \nu)}{q(\theta)} d\theta \\ & \quad - \int q(\theta) \ln \frac{p(\theta, \mathbf{X}|m_0, \rho_0, \mathbf{W}_0, \eta_0, \nu)}{q(\theta)} d\theta \end{aligned} \quad (10)$$

where $\mathbf{X}=\{\mathbf{x}_n\}$, $\theta = \{\boldsymbol{\mu}, \boldsymbol{\Lambda}, \mathbf{u}\}$, $\mathbf{u} = \{u_n\}$, and $q(\theta)$ is the so-called variational distribution over $\boldsymbol{\mu}, \boldsymbol{\Lambda}$, and \mathbf{u} , such that $q(\theta) = q(\boldsymbol{\mu})q(\boldsymbol{\Lambda})q(\mathbf{u})$ (assuming $q(\boldsymbol{\mu})$, $q(\boldsymbol{\Lambda})$, and $q(\mathbf{u})$ are independent).

The second term of Equation (10) is the Kullback-Leibler (KL) divergence between $q(\theta)$ and $p(\theta|\{x_n\}, m_0, \rho_0, \mathbf{W}_0, \eta_0, \nu)$, which is non-negative and zero only if the two distributions are identical. Thus, the first term can be understood as the lower bound of the log-marginal likelihood $\Lambda(q)$. Therefore, seeking to minimize the second term of Equation (10) amounts to maximizing the lower bound $\Lambda(q)$.

Accordingly, one way to represent the lower bound $\Lambda(q)$ is shown:

$$L(q) \equiv \int q(\theta) \ln \left\{ \frac{p(\mathbf{X}, \theta)}{q(\theta)} \right\} d\theta \leq \ln p(\mathbf{X}) \quad (11)$$

where θ represents the set of all unobserved stochastic variables.

In Equation (11), $q(\theta)$ represents the variational posterior distribution, and $p(\mathbf{X}, \theta)$ is the joint distribution over the stochastic modeling parameters. The difference between the right hand side of Equation (11) and $\Lambda(q)$ is given by the

KL divergence $KL(q||p)$ between the variational posterior distribution $q(\theta)$ and the true posterior distribution $p(\theta, \mathbf{X})$.

Given the priors of Equations (5), (6), (7), and (8), the variational posterior distributions $q(\cdot)$ for \mathbf{s} , π , μ_m , Λ_m , and u may be computed.

For $q(\mathbf{s})$, where \mathbf{s} represents the labeling parameters:

$$q(\mathbf{s}) = \prod_{n,m}^{N,M} p_{nm}^{s_{nm}} \quad (12)$$

where

$$p_{nm} = \frac{r_{nm}}{\sum_{m=1}^M r_{nm'}} \quad (13)$$

where, in turn,

$$r_{nm} = \exp \left(\langle \ln \pi_m \rangle + \frac{1}{2} \langle \ln |\Lambda_m| \rangle + \frac{d}{2} \langle \ln u_{nm} \rangle - \frac{\langle u_{nm} \rangle \langle \Delta_{nm}^2 \rangle}{2} - \frac{d}{2} \ln 2\pi \right) \quad (14)$$

Although the last term in the argument for the exponential cancels out in Equation (13). In addition,

$$\langle \ln |\Lambda_m| \rangle = d \ln 2 - \ln |\mathbf{W}| + \sum_{i=1}^d \Psi \left(\frac{\eta + 1 - i}{2} \right) \quad (15)$$

$$\langle \Delta_n^2 \rangle = \mathbf{x}_n^\top \eta_m \mathbf{W}_m \mathbf{x}_n - 2 \mathbf{x}_n^\top \eta_m \mathbf{W}_m \mathbf{m}_m + \text{Tr} \left[(\mathbf{m}_m \mathbf{m}_m^\top + R_m^{-1}) \eta_m \mathbf{W}_m \right] \quad (16)$$

and

$$\langle s_{nm} \rangle = p_{nm} \quad (17)$$

For $q(\pi)$, where π represents the mixing coefficients:

$$q(\pi) = D(\pi | \alpha) \quad (18)$$

where

$$\alpha_m = \sum_{n=1}^N \langle s_{nm} \rangle + \hat{\alpha}_m \quad (19)$$

and

$$\langle \pi_m \rangle = \frac{\alpha_m}{\alpha_0} \quad (20)$$

where $\alpha_0 = \sum_{m'} \alpha_{m'}$ and $m'=1, \dots, M$. Furthermore, $\langle \ln \pi_m \rangle = \Psi(\alpha_m) - \Psi(\alpha_0)$, where

$$\Psi(a) = \frac{d \ln \Gamma(a)}{da} \quad (21)$$

For $q(\mu_m)$, where μ_m represents the mean of the m^{th} Student distribution component in the mixture:

$$q(\mu_m) = N(\mu_m | m_m, \mathbf{R}_m) \quad (22)$$

where

$$\mathbf{R}_m = \langle \Lambda_m \rangle \sum_{n=1}^N \langle w_{nm} \rangle + \rho_0 \mathbf{I} \quad (23)$$

$$m_m = \mathbf{R}_m^{-1} \left(\langle \Lambda_m \rangle \sum_{n=1}^N \langle w_{nm} \rangle \mathbf{x}_n + \rho_0 m_0 \right)$$

and

$$\langle w_{nm} \rangle = \langle s_{nm} \rangle \langle u_{nm} \rangle \quad (24)$$

For $q(\Lambda_m)$, where Λ_m represents the precision matrix of the m^{th} Student distribution component in the mixture:

$$q(\Lambda_m) = \mathcal{W}(\Lambda_m | \mathbf{W}_m, \eta_m) \quad (25)$$

where

$$\mathbf{W}_m^{-1} = \mathbf{W}_0^{-1} + \sum_n \langle w_{nm} \rangle (\mathbf{x}_n \mathbf{x}_n^T - \mathbf{x}_n m_m^T - m_m \mathbf{x}_n^T + (m_m m_m^T + R_m^{-1})) \quad (26)$$

and

$$\eta_m = \eta_0 + \hat{s}_m \quad (27)$$

where $\hat{s}_m = \sum_n \langle s_{nm} \rangle$.

For $q(\mathbf{u})$, where \mathbf{u} represents the scaling parameters of the precision matrices:

$$q(u_{nm}) = G(u_{nm} | a_{nm}, b_{nm}) \quad (28)$$

where

$$a_{nm} = \frac{\nu_m + \langle s_{nm} \rangle d}{2} \quad (29)$$

where d represents the dimensionality of the data,

$$b_{nm} = \frac{\nu_m + \langle s_{nm} \rangle \langle \Delta_{nm}^2 \rangle}{2} \quad (30)$$

and

$$\langle \Delta_{nm}^2 \rangle = \mathbf{x}_n^\top \boldsymbol{\eta}_m \mathbf{W}_m \mathbf{x}_n - 2 \mathbf{x}_n^\top \boldsymbol{\eta}_m \mathbf{W}_m \mathbf{m}_m + \text{Tr} \left[(\mathbf{m}_m \mathbf{m}_m^\top + \mathbf{R}_m^{-1}) \boldsymbol{\eta}_m \mathbf{W}_m \right] \quad (31)$$

A constrained family of distributions for $q(\theta)$ is chosen such that the lower bound $\Lambda(q)$ becomes tractable. The optimal member of the family can then be determined by maximization of $\Lambda(q)$, which is equivalent to minimization of the KL divergence. Thus, the resulting optimal solution for $q(\theta)$ represents an approximation of the true posterior of $p(\theta | \{x_n\}, \mathbf{m}_0, \rho_0, \mathbf{W}_0, \eta_0, \nu)$, assuming a factorized variational distribution for $q(\theta)$ of:

$$q(\theta) = q(\{\boldsymbol{\mu}_m\}) q(\{\boldsymbol{\Lambda}_m\}) q(\pi) q(\{\mathbf{s}_n\}) q(\{\mathbf{u}_n\}) \quad (32)$$

A free-form variational optimization is now possible with respect to each of the individual variational factors of Equation (32). Because the variational factors are coupled, the variational approximations of the factors are computed iteratively by first initializing the distributions, and then cycling to each factor in turn and replacing its current estimate by its optimal solution, given the current estimates for the other factors, to give a new approximation of $q(\theta)$. Interleaved with the optimization with respect to each of the individual variational factors, the lower bound is optimized with respect to each of the non-stochastic parameters ν_m by employing standard non-linear optimization techniques. The lower bound $\Lambda(q)$ is then computed using the new approximation of $q(\theta)$ for the current iteration

In one implementation, the iteration continues until the lower bound $\Lambda(q)$ changes by less than a given threshold. In an alternative implementation, $q(\theta)$ may also be tested prior to computation of the lower bound $\Lambda(q)$ in each iteration,

1 such that if the value of $q(\theta)$ changes by less than another given threshold, then the
2 iteration skips the computation and testing of the lower bound $\Lambda(q)$ and exits the
3 loop. In yet another implementation, individual factors of Equation (32) may be
4 tested to determine whether to terminate the optimization of the modeling
5 parameters.

6 In the described approach, approximate posterior distributions of the
7 stochastic modeling parameters $\{\mu_m, \Lambda_m\}$, π , and $\{s_m, u_n\}$, as well as a value of the
8 modeling parameter ν , are determined. Given these modeling parameters, the
9 Student mixture density of Equation (3) can be obtained to model the input data.

10 FIG. 2 illustrates exemplary operations 200 for robust Bayesian mixture
11 modeling. A receiving operation 202 receives prior distributions of each modeling
12 parameter in the set of modeling parameters for a mixture of Student distributions.
13 In one implementation, the prior distributions may be computed using the
14 Equations (5), (6), (7), and (8), although other prior distributions may be used in
15 alternative embodiments. As such, an operation of computing the prior
16 distributions (not shown) may also be included in an alternative implementation.

17 Another receiving operation 204 receives the independent, identically
18 distributed data. Exemplary data may include without limitation auditory speech
19 data from an unknown number of speakers, where determining the correct number
20 of speakers is part of the modeling process and image segmentation data from
21 images containing few large and relatively homogeneous regions as well as
22 several very small regions of different characteristics (outlier regions), where
23 modeling of the few larger regions should not be notably affected by the presence
24 of the outlier regions.

1 Yet another receiving operation 206 receives initial estimates of the
2 posterior distributions for a set of modeling parameters for a mixture of Student
3 distributions. The initial estimates may be received from another process or be
4 determined in a determining operation (not shown) using a variety of methods,
5 including a random approach. However, the optimization of the modeling
6 parameter can resolve quicker if the initial estimates are closer to the actual
7 posterior distributions. In one implementation, heuristics are applied to the prior
8 distributions to determine these initial estimates. In a simple example, the
9 posteriors are set equal to the priors. A more elaborate example is to heuristically
10 combine the priors with the results of fast, non-probabilistic methods, such as *K*-
11 means clustering.

12 A selection operation 208 selects one of the modeling parameters in the set
13 of modeling parameters. A computation operation 210 computes a tractable
14 approximation of the posterior distribution of the selected modeling parameter
15 using the current estimates of the other modeling parameters. (In the first
16 iteration, the current estimates of the other modeling parameters represent their
17 initial estimates.) In one implementation, the current state of the estimate of each
18 modeling parameter is stored in a storage location, such as in a memory.

19 In the illustrated implementation, a variational inference method produces
20 the tractable approximation. In one variational inference approach, the tractable
21 posterior distribution is approximated using the Equations (12), (18), (22), (25),
22 and (28). The tractable approximation of the selected modeling parameter
23 becomes the current estimate of that modeling parameter, which can be used in
24
25

1 subsequent iterations. Alternatively, other approximation methods, including
2 Monte Carlo techniques, may be employed.

3 A computation operation 212 computes the lower bound of the log
4 marginal likelihood, such as by using Equation (11). If the lower bound is
5 insufficiently optimized according to the computation operation 212, such as by
6 improving by greater than a given threshold or by some other criterion, a decision
7 operation 214 loops processing back to the selection operation 208, which selects
8 another modeling parameter and repeats operation 210 212 and 214 in a
9 subsequent iteration. However, if the lower bound is sufficiently optimized,
10 processing proceeds to a generation operation 216, which generates the probability
11 density of the data based on the mixture of Student distributions characterized by
12 the current estimates of the modeling parameters (e.g., using Equation (4)).

13 It should be understood that the order of at least some of the operations in
14 the described process may be altered without altering the results. Furthermore,
15 other methods of determining whether the posterior distribution approximations of
16 the modeling parameters are satisfactorily optimized, including testing whether the
17 individual posterior distribution factors (e.g., $q(s)$) change little in each iteration or
18 testing whether the product (e.g., $q(\theta)$) of the posterior distribution factors changes
19 little in each iteration.

20 FIG. 3 illustrates an exemplary robust Bayesian mixture modeling
21 system 300. Inputs to the system 300 include input data 302, initial estimates of
22 the modeling parameters 304, and prior distributions of the modeling
23 parameters 306.

1 A modeling parameter selector 308 selects a modeling parameter that is to
2 be approximated in each iteration. A tractable approximation module 310 receives
3 the inputs and the selection of the modeling parameter to generate a tractable
4 approximation of the selected modeling parameter (e.g., based on variational
5 inference or Monte Carlo techniques). In one implementation, the tractable
6 approximation module 301 also maintains a current state of the estimate of each
7 modeling parameter in a storage location, such as in a memory.

8 Based on the current estimates of the modeling parameters, including the
9 new approximation of the selected modeling parameter, a lower bound optimizer
10 module 312 computes the lower bound of the log marginal likelihood. If the lower
11 bound fails to satisfy an optimization criterion (such as by increasing more than a
12 threshold amount), the lower bound optimizer module 312 triggers the modeling
13 parameter selector module 308 to select another modeling parameter in a next
14 iteration. Otherwise, the current estimates of the modeling parameters are passed
15 to a data model generator 314, which generates a data model 316 including the
16 probability density of the data based on the mixture of Student distributions
17 characterized by the current estimates of the modeling parameters (e.g., using
18 Equation (4))

19 The exemplary hardware and operating environment of FIG. 4 for
20 implementing the invention includes a general purpose computing device in the
21 form of a computer 20, including a processing unit 21, a system memory 22, and a
22 system bus 23 that operatively couples various system components include the
23 system memory to the processing unit 21. There may be only one or there may be
24 more than one processing unit 21, such that the processor of computer 20
25

1 comprises a single central-processing unit (CPU), or a plurality of processing
2 units, commonly referred to as a parallel processing environment. The computer
3 20 may be a conventional computer, a distributed computer, or any other type of
4 computer; the invention is not so limited.

5 The system bus 23 may be any of several types of bus structures including a
6 memory bus or memory controller, a peripheral bus, a switched fabric, point-to-
7 point connections, and a local bus using any of a variety of bus architectures. The
8 system memory may also be referred to as simply the memory, and includes read
9 only memory (ROM) 24 and random access memory (RAM) 25. A basic
10 input/output system (BIOS) 26, containing the basic routines that help to transfer
11 information between elements within the computer 20, such as during start-up, is
12 stored in ROM 24. The computer 20 further includes a hard disk drive 27 for
13 reading from and writing to a hard disk, not shown, a magnetic disk drive 28 for
14 reading from or writing to a removable magnetic disk 29, and an optical disk drive
15 30 for reading from or writing to a removable optical disk 31 such as a CD ROM
16 or other optical media.

17 The hard disk drive 27, magnetic disk drive 28, and optical disk drive 30
18 are connected to the system bus 23 by a hard disk drive interface 32, a magnetic
19 disk drive interface 33, and an optical disk drive interface 34, respectively. The
20 drives and their associated computer-readable media provide nonvolatile storage
21 of computer-readable instructions, data structures, program modules and other
22 data for the computer 20. It should be appreciated by those skilled in the art that
23 any type of computer-readable media which can store data that is accessible by a
24 computer, such as magnetic cassettes, flash memory cards, digital video disks,
25

1 random access memories (RAMs), read only memories (ROMs), and the like, may
2 be used in the exemplary operating environment.

3 A number of program modules may be stored on the hard disk, magnetic
4 disk 29, optical disk 31, ROM 24, or RAM 25, including an operating system 35,
5 one or more application programs 36, other program modules 37, and program
6 data 38. A user may enter commands and information into the personal
7 computer 20 through input devices such as a keyboard 40 and pointing device 42.
8 Other input devices (not shown) may include a microphone, joystick, game pad,
9 satellite dish, scanner, or the like. These and other input devices are often
10 connected to the processing unit 21 through a serial port interface 46 that is
11 coupled to the system bus, but may be connected by other interfaces, such as a
12 parallel port, game port, or a universal serial bus (USB). A monitor 47 or other
13 type of display device is also connected to the system bus 23 via an interface, such
14 as a video adapter 48. In addition to the monitor, computers typically include
15 other peripheral output devices (not shown), such as speakers and printers.

16 The computer 20 may operate in a networked environment using logical
17 connections to one or more remote computers, such as remote computer 49. These
18 logical connections are achieved by a communication device coupled to or a part
19 of the computer 20; the invention is not limited to a particular type of
20 communications device. The remote computer 49 may be another computer, a
21 server, a router, a network PC, a client, a peer device or other common network
22 node, and typically includes many or all of the elements described above relative
23 to the computer 20, although only a memory storage device 50 has been illustrated
24 in FIG. 4. The logical connections depicted in FIG. 4 include a local-area network
25

1 (LAN) 51 and a wide-area network (WAN) 52. Such networking environments
2 are commonplace in office networks, enterprise-wide computer networks, intranets
3 and the Internet, which are all types of networks.

4 When used in a LAN-networking environment, the computer 20 is
5 connected to the local network 51 through a network interface or adapter 53,
6 which is one type of communications device. When used in a WAN-networking
7 environment, the computer 20 typically includes a modem 54, a network adapter, a
8 type of communications device, or any other type of communications device for
9 establishing communications over the wide area network 52. The modem 54,
10 which may be internal or external, is connected to the system bus 23 via the serial
11 port interface 46. In a networked environment, program modules depicted relative
12 to the personal computer 20, or portions thereof, may be stored in the remote
13 memory storage device. It is appreciated that the network connections shown are
14 exemplary and other means of and communications devices for establishing a
15 communications link between the computers may be used.

16 In an exemplary implementation, a modeling parameter selector, a tractable
17 approximation module, a lower bound optimizer module, a data model generator,
18 and other modules may be incorporated as part of the operating system 35,
19 application programs 36, or other program modules 37. Initial modeling
20 parameter estimates, input data, modeling parameter priors, and other data may be
21 stored as program data 38.

22 The embodiments of the invention described herein are implemented as
23 logical steps in one or more computer systems. The logical operations of the
24 present invention are implemented (1) as a sequence of processor-implemented
25

1 steps executing in one or more computer systems and (2) as interconnected
2 machine modules within one or more computer systems. The implementation is a
3 matter of choice, dependent on the performance requirements of the computer
4 system implementing the invention. Accordingly, the logical operations making
5 up the embodiments of the invention described herein are referred to variously as
6 operations, steps, objects, or modules.

7 The above specification, examples and data provide a complete description
8 of the structure and use of exemplary embodiments of the invention. Since many
9 embodiments of the invention can be made without departing from the spirit and
10 scope of the invention, the invention resides in the claims hereinafter appended.